2P-3

ファイル利用履歴から抽出した抽象タスク間ワークフローに基づくファイルと操作の推薦

宋 強† 川端 貴幸†† 伊藤 史朗†† 渡辺 陽介††† 横田 治夫†

† 東京工業大学大学院情報理工学研究科計算工学専攻
†† キヤノン（株）ソフトウェア応用技術開発センター
††† 東京工業大学学術国際情報センター

1 はじめに

近年ファイルシステム内のファイル数が爆発的に増加している．作業を達成するために必要な文書の検索時間や作業の流れの習得にかかる労力を減らすために，大量ファイルの中から適切なファイルを，操作手順と伴に推薦することで，利用者の支援が可能である．従来の協調フィルタリングなどのファイル推薦技術では操作の順番に厳しすぎる上，過去に使用されたファイルしか対象にできなかった．そこで我々の研究グループは，過去に利用されていないファイルも対象とし，操作の順番が異なっても推薦できる方法を提案してきた[1]．そのために，ファイル利用履歴から，ファイル操作のまとまりである抽象タスクを求め，更に抽象タスク間の頻出抽象ワークフローを抽出し，記憶しておく．利用者のその時点でのファイル操作パターンに対応する抽象ワークフローを検出し，推薦を行う．ユーザの特定のファイルに対する操作としてではなく，同じ特徴でグループ化されたファイルクラスタへの操作と見なすことで，新規ファイルに対しても推薦可能になった．本稿では，[1] の上で，ファイルの抽象化のための新たなファイル間類似度の定義方法を追加して，比較実験を行った．

2 提案手法

図 1 で示すように，ユーザがファイルにアクセスすると，ファイルアクセス履歴が残る．我々はこのファイルアクセス履歴から，操作パターンなどの情報を抽出してデータベースに保存する．また，ユーザのファイル操作を監視し，ログから抽出したワークフローなどと照合しながら，ユーザの直前の操作によって次にアクセスする可能性が高いファイル及び操作を推薦する．システムはオフライン部とオンライン部の二つの部分から構成される．この章で，システムの内部処理について説明する．

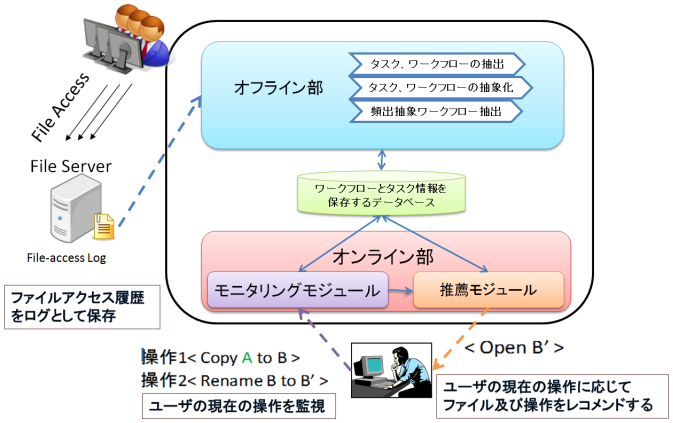

図 1: 提案手法の全体図

2.1 オフライン部

[**タスクとワークフローの抽出**] まず，ユーザごとにファイルアクセスログを分割する．ユーザごとのログ内で前後二記録の間のアクセス時間差（パラメータ：タスクギャップタイム）が一定時間以上であれば，1 つのタスクが終了し，新たなタスクが開始されたものとみなし，タスクを切り出す．また，もっと長い一定時間差（パラメータ：ワークフローギャップタイム）が見られた場合には，新たなワークフローの開始とみなす．

[**タスクとワークフローの抽象化**] ファイル名の一部が異なっても，性質的に同じグループ（例えば，議事録ファイルグループ）のファイル同士を纏めるために，ファイルをグループ化する．そのために，ファイル間類似度を定義する必要がある．ファイル間類似度をコピー関係類似度とファイル名類似度の二つの面から定

Recommendation Method for Files and Operations based on Workflows of Abstract Tasks from Access Log

Qiang SONG†, Takayuki KAWABATA†† Fumiaki ITOH††
Yousuke WATANABE††† and Haruo YOKOTA†
†Dept. of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology
††Canon Inc. Applied Software Technology Development Center
†††Global Scientific Information and Computing Center, Tokyo Institute of Technology
soukyou@de.cs.titech.ac.jp kawabata.takayuki@canon.co.jp
ito.fumiaki@canon.co.jp watanabe@de.cs.titech.ac.jp
yokota@cs.titech.ac.jp

Copyright ©2013 Information Processing Society of Japan.
All Rights Reserved.

義する．コピー関係類似度とは，コピー関係を持つファイル同士に与える類似度である．ファイル名類似度については，[1] ではファイル名をキャラクタのシーケンスとして，お互いのシーケンスの類似度合を最長共通サブシーケンス (LCS)[2] を計算することで定義した．その上で，我々は今回キャラクタレベルの編集距離 (LD)，キャラクタレベルの 2-gram を求めることでの類似度算出方法を追加する．それに，ファイル名をキャラクタそのままのシーケンスとするのではなく，品詞に区切ってから，LCS や LD を求めることでの類似度算出方法も提案する．

次に，ファイルを抽象化する．具体的に，ファイル間類似度によってファイルをクラスタリングし，タスク内に出現するファイルをファイルクラスタに置き換える．更に，ファイル間の類似度を用いて，タスク間同士の類似度を計算して，類似度が高いタスク群をクラスタとしてまとめて，タスクの抽象化を行う．こうすることで，抽象ワークフローを抽象タスクのシーケンスとして抽出する．

[**頻出抽象ワークフローの抽出**] 作成された抽象ワークフローの集合から，出現頻度が低いワークフローを除くために，シーケンシャルパターンマイニングで頻出するシーケンスを抽出する．

2.2 オンライン部

[**モニタリングモジュール**] モニタリングモジュールの役割はユーザの現在のファイル操作情報からユーザが行なっているワークフローの推定である．同じ作業のパターンと言っても，扱うファイルが異なると考えられるため，ユーザが特定のファイルに対する操作としてではなく，同じ特徴でグループ化されたファイルクラスタへの操作と見なす．それから，抽象タスクと頻出抽象ワークフローの推定を行う．ここで，頻出抽象ワークフローを選ぶ基準として，ユーザが現在操作しているワークフローとの一致度合, 頻出抽象ワークフローの出現頻度と頻出抽象ワークフローの推薦可能要素数の三つの基準を考慮する．

[**推薦モジュール**] 推薦モジュールの役割は推定した頻出抽象ワークフローに基づき，次にユーザが操作する可能性が高いファイルと操作の推薦を行うことである．ユーザが現在行っているタスクと，そのタスクの属しているワークフローがわかれば，そこから次の抽象タスクが特定できる．

頻出抽象ワークフローに基づいて次に作業される対象のファイルクラスタを特定できた後に，クラスタ内にある適切な実際のファイルをピックアップしてユーザに推薦することになる．しかし，各ファイルクラスタは複数個のファイルを含むため，推薦の際には複数の候補が存在する．ここでは，過去に一番アクセスされた回数が多いファイルを優先的に推薦するものとする．

3 評価実験

実験の目的は，提案手法の特徴である抽象化とタスク導入による有効性を検証することと，抽象化の処理

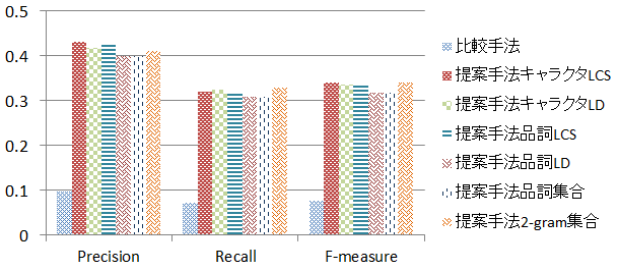

図 2: 提案手法と比較手法の比較

におけるファイル間類似度算出方法の比較である．

比較手法として，抽象タスクなしで，ファイル操作のシーケンスから直接頻出シーケンスを抽出し，それをワークフローとする方法を用いる．今回の実験では，実際に企業から提供された実ファイルアクセスログを用いて行った．結果を図 2 で示す．比較手法と比べ，提案手法の方の Precision, Recall と F-measure が向上したことが確認でき，抽象化とタスクの概念の導入の有効性が確認できた．

異なる類似度算出方法同士の比較として，もし Recall 重視なら，2-gram の方法が一番適していることが分かった．F-measure で総合的に見ると，キャラクタレベルでの LCS か 2-gram が一番良かったことが分かった．また，ファイル名を単語に分けた上での抽象化は Recall などが落ちったことも分かった．これの原因としては，ファイル名を単語レベルに分けると，ファイル名同士の共通部分の判定が単語ごとに行われることで，基準が厳しくなり，高い類似度が得られにくくなったためと考えられる．

4 まとめと今後の課題

本論文では，アクセスログに基づくファイル推薦においてファイルを抽象化するための，ファイル同士のファイル名による類似度計算方法を新たに提案した．それに，評価実験と通して比較した．

これからの課題として，Recall の向上，他のログでの実験などが挙げられる．

参考文献

[1] 宋強, 川端貴幸, 伊藤史朗, 渡辺陽介, 横田治夫, “ファイルレコメンデーションのためのファイル利用履歴に基づくタスク間ワークフロー抽出手法”, 第四回データ工学と情報マネジメントに関するフォーラム,2012

[2] Daniel S. Hirschberg, “Algorithms for the Longest Common Subsequence Problem”, Journal of the ACM (JACM) Vol 24 Issue 4, 1977.

Copyright ©2013 Information Processing Society of Japan.
All Rights Reserved.